Sādhanā (2020)45:254
https://doi.org/10.1007/s12046-020-01494-z

An automated approach to retrieve lecture videos using context based semantic features and deep learning

N POORNIMA[1,2] and B SALEENA[1,*]

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India
[2]School of Computing, SRM Institute of Science and Technology, Chennai, India
e-mail: poornimn1@srmist.edu.in; saleena.b@vit.ac.in

MS received 17 February 2020; revised 9 July 2020; accepted 17 August 2020

Abstract. Video digitization is one of the emerging technologies holding significant importance over years in applications like video recording and video compression. There are different techniques available in the literature for the effective retrieval of videos. This research work presents a video retrieval scheme based on a deep learning strategy. Initially, the video archive is subjected to the keyframe extraction, for extracting useful keyframes from the video. The features extracted from the keyframes are stored in the feature database. The features are clustered using the Fuzzy C Means (FCM) algorithm. These clustered features have been provided to the deep learner for finding the optimal centroid for the incoming user query. For experimentation, the research has considered videos from different categories, and both the text query and the video query have been used for the retrieval. The experimental results demonstrate the efficiency of the proposed deep learning strategy in video retrieval and its achievement of improved values of 0.9620, 0.9682, and 0.9652 respectively for recall, precision, and F-measure.

Keywords. Video retrieval; keyframes; clustering; deep learning.

1. Introduction

Digitization of the lecture contents by recording it in the form of videos has helped universities and colleges in the improvement of teaching skills [1]. Students prefer learning materials that are in the video format as they are easily available on online platforms. Thus, the lecture videos have improved the robustness of the study material [2]. Video is the combination of the text, image, and sound, and hence, the use of the lecture videos as the study material enables a live study experience for the students [3]. Also, universities can upload the study material in their portals to make it readily available for the students. Using lecture videos as study material has been in vogue in recent years in most of the universities. Some colleges record the presentation of the lecturer and upload it in the internet platform [4]. Direct recording of the presentations may increase the multimedia content on the internet and it is extremely difficult for the students to find the actual content from a large source of information [1]. The ever increasing demand for lecture videos has given rise to a video retrieval system that analyzes the large database and retrieves similar video content related to the query issued by the user. As the video archives on the internet are very large, retrieval of similar video contents is a complex task [5, 6].

Most of the video retrieval schemes use the search function for retrieving the video contents. As the size of the video database is large, retrieving similar videos without the search function is nearly impossible. Also, it is difficult for the user to ascertain the correctness of the retrieved video contents without opening the video file [7]. Sometimes, the video may cover only a very little portion related to the query rather than providing any additional information, which makes the process of retrieval unsuccessful. Hence, it is necessary to build a video retrieval system for the user with appropriate file contents [8, 9]. Video search engines, such as YouTube, Bing, and Vimeo, reply with the video files based on the title, genre, person, etc. Most of the metadata information is created manually by the user and may mislead the contents sometimes. Hence, the video retrieval system should automatically generate the metadata based on the contents, to improve the quality of the retrieval process [1]. Literature has provided two different schemes for retrieving the contents of the video. They are described as manual and automatic schemes. The manual approach is considered to be more accurate than the automatic scheme but requires a longer time and cost for the retrieval process [10]. The automatic scheme uses the low-level video analysis approach for analyzing the contents of the video [11].

The video file is a combination of several text files, audios, and images, and hence, there is a need for video retrieval to extract the related feature contents for video

*For correspondence
Published online: 08 October 2020

retrieval [12]. The video retrieval scheme extracts the features from the video through several techniques, such as Text-based, Audio-based, Metadata-based, and Content-based techniques [13]. Use of metadata type feature extraction helps retrieval of information relating to the type, the title, date, etc., from the video. Meanwhile, the text-based technique extracts the text available in the video through the Optical character recognition (OCR) based scheme. The audio based scheme uses different speech recognition based techniques in the feature extraction for retrieving the audio contents of the video file. The content-based feature extraction technique is said to be a combination of all the above mentioned techniques [14]. Content Based Video Retrieval (CBVR) is considered to be the most successful technique for retrieving videos from large video archives. The CBVR technique retrieves the lecture video contents with a smaller number of keywords. The term 'content' in the CBVR may refer to color, texture, text, or audio. Also, the CBVR technique responds to the image query provided by the user. Despite the CBVR technique having proved its utility for retrieval of the digital video contents, an increase in the volume of media content on the internet has made the retrieval process a complete task. Using more digital libraries or repositories may improve the video retrieval process [15]. In [16], On-the-fly Video Retrieval has been presented for retrieving the video contents.

This paper proposes a video retrieval strategy using the deep learning based scheme. Initially, the keyframes from the input video frames are generated, and then, the feature database is constructed by extracting the keywords, semantic words, contextual features, together with the image texture, which is extracted using the Local Directional Pattern (LDP) [17]. The features extracted are clustered using Fuzzy C Means (FCM) for the indexing, and are used for training the deep learner with respect to the relevant clusters. Finally, the features are given as input to the trained deep learning for the output query to find the relevant cluster of relevant videos.

The major highlights of this research work are the design and development of the deep learning-based video retrieval strategy for the retrieval of lecture videos from a large database. This works specifically extracts the contextual features along with other features for retrieval purposes.

The remaining sections of the paper are organized as follows: The video retrieval strategy and the various techniques used for the retrieval process is discussed in section 1. Section 2 reviews some of the existing works that have contributed to the video retrieval through classification and clustering approach. Section 3 discusses the newly proposed deep learning based video retrieval technique. Section 4 highlights the experimental results achieved by using the proposed system and also presents a comparative analysis of the performance of the proposed video retrieval technique with the other techniques. The conclusion and scope for future enhancements are discussed in section 5.

2. Literature review

An automatic video indexing approach for retrieving video content was developed by Haojin Yang and Christoph Meinel [18]. They have adopted techniques, such as OCR and automatic speech recognition, for retrieving the features from the database. The feature extraction extracts useful keyframes for the retrieval process. Even though the technique has improved performance, the presence of noise reduces the overall performance. Kai Li *et al* [19] have presented an automated video retrieval system for capturing and detecting similar videos in the video archive. The retrieval was done by analyzing the text contents and the keywords in the video. The system yielded reduced performance while using multi-videos. Esha Baidya and Sanjay Goel [20] have proposed an automated video retrieval scheme using the OCR technique. Using the OCR, the important information was extracted from the video, and further, the technique collected the embedded information in the video slides. But the performance is degraded with high recall value. Nhu Van Nguyen *et al* [21] have presented a video retrieval scheme with document analysis. The scheme has adopted the graphical text localization and recognition techniques for extracting the keywords from the video. The scheme performed well while using the multi-modal and cross-modal videos. The system exhibited errors in retrieval.

Araujo and Girod [22] have proposed an asymmetric comparison technique for video retrieval. The scheme explored the database by incorporating Fisher vectors. The technique works in a flexible retrieval environment. Rahmani and Zargari [23] have proposed a feature vector for the video retrieval process which involves an analysis of the motion structure of the video sequences. Besides improved results, the scheme faces complexity issues during the analysis of large video contents. Lin *et al* [24] presented the deep learned global descriptors for video retrieval. The deep learned global descriptors depend on the invariance theory. The authors have further proposed the Nested Invariance Pooling (NIP) scheme for analyzing the pooled descriptors in the video. The scheme has complementary effects on the handcrafted descriptors. Rouhi and Thom [25] have proposed the CBVR scheme using different encoding profiles. The scheme has analyzed the effects of the encoding scheme during the retrieval process. The scheme provided improved tolerance and robustness towards the noise and different encoding types.

2.1 *Challenges*

Challenges in developing the retrieval system for lecture videos are as follows:

- The critical challenge is the recognition of the teaching topic by the retrieval system. Unavailability of the

teaching topic may increase the complexity of the retrieval process [3].

- The lecture video files have low level correlation among the features of different videos, and hence, it is more challenging to retrieve the lecture video compared to other video files [3].
- Some works have adopted the CNN based global descriptors for video retrieval, but they face many challenges. The initial challenge in adopting the CNN based method for video retrieval is the absence of the invariance in the CNN method, while geometric transformations occur in the input image. The geometric variations in the image include the rotation of the image in the consecutive frame [7].
- Another challenge confronting the CNN based method for video retrieval is the performance degradation for the rotated image query due to the global descriptors [7].
- The use of the conventional manual video descriptors has reduced the performance of the video retrieval technique [7] as they are robust towards the scale and rotation changes occurring in the 2D plane.

3. Proposed video retrieval scheme using deep learning

A video retrieval scheme with a deep learning strategy is explained in this section. Lecture video retrieval designed in this work includes feature extraction, clustering, and deep learning. The proposed architecture for video retrieval using deep learning techniques is depicted in figure 1.

Keyframes are extracted from the videos present in the database. After extracting the keyframes from the video, the features, such as words, semantic words, context words, and LDP features, are extracted and the feature database is created. Then, the database is subjected to the FCM clustering and finally, DBN that gets trained with the cluster centroids is used for the retrieval of the video. While the user gives a search query arrives in the video retrieval system, the above mentioned features are extracted from the query and given to the DBN for testing. The DBN classifier identifies the optimal cluster belonging to the query, and the videos related to the optimal cluster are retrieved by the proposed deep learning based video retrieval scheme.

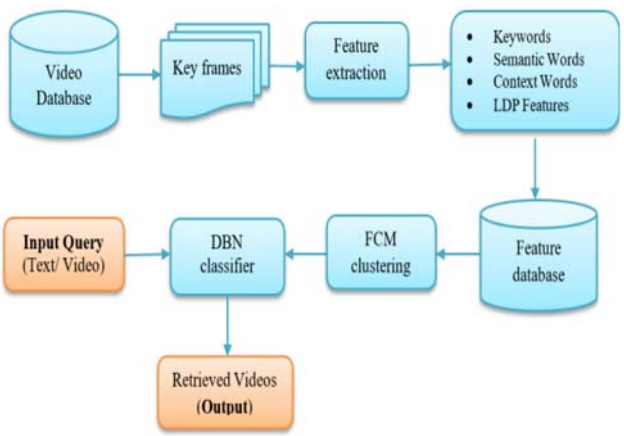

Figure 1. The proposed architecture for video retrieval using deep learning.

3.1 *Extraction of the Keyframe from an input video*

The initial stage in the proposed deep learning based video retrieval system is the extraction of the keyframes from the video. To extract important features from the video, a video must be broken down into atomic units called frames. Frames may repeat in the video for visibility. To avoid redundancy, unique frames in the video are identified which are called keyframes. For content based processing of video, video structure analysis is carried out by dividing the video into substantial components which include scene separation, shot boundary detection, and keyframe extraction. Since redundancy is the most important property of a video, Keyframe extraction plays a vital role in video structure analysis. The redundant frames are excluded from the videos for lesser processing and to make the video more compact and useful. Keyframe extraction is the method of excerpting a frame or set of frames which summarizes the video by covering all the major events of the video.

Slides in the lecture video provide an outline for understanding the video. Therefore, each unique slide from lecture videos is considered as a keyframe. After fragmenting a video into a set of keyframes, the text extraction procedure is applied to each keyframe to detect and recognize the textual information.

For the experimentation, this work considers the video archive D with V number of lecture video contents, expressed as,

$$D = \{R_1, R_2, \ldots, R_i, \ldots, R_V\} \tag{1}$$

Where R_i represents the i^{th} lecture video in the video archive and V refers to the total lecture videos in the database. The lecture videos have substantial information for processing. The use of all the data in the video for processing makes the video retrieval more complex. Hence, this work extracts some of the keyframes in the video for the analysis. Consider the video R_i having K number of keyframes, and then after the keyframe extraction, the video R_i is represented as,

$$R_i = \left\{R_i^1, R_i^2, \ldots, R_i^k, \ldots, R_i^K\right\} \tag{2}$$

where R_i^k represents the k^{th} keyframe in the video R_i . Keyframe extraction improves the video retrieval process

as most of the important image textures and features are present in the keyframe.

3.2 *Feature extraction*

The steps involved in the construction of feature databases include keyword extraction, semantic word identification, contextual word detection, and image texture extraction. Various features extracted from the keyframes are shown in figure 2.

3.2a *Extracting the keywords using the OCR technique*: The keyframes in the video contain several keywords useful for video retrieval. Here, the OCR [26] is used for retrieving the keywords from the keyframes. Consider there are w numbers of keywords on a keyframe and the extracted keywords through the OCR technique are represented as

$$s = \{s_1, s_2, \ldots, s_z, \ldots, s_w\} \tag{3}$$

where s_z indicates the z^{th} keyword in the keyframe.

3.2b *Semantic words*: Semantic words are extracted from the keywords after the identification of these keywords. The semantic words are the keywords having a similar synonym and they are expressed as,

$$L = \{L_1^1, L_2^2, \ldots, L_w^V\} \tag{4}$$

where L_1^1 indicates the semantic words for the keyword s_1 and the semantic words are expressed as follows,

$$L_1^1 = \{s_{w1}^1, s_{w2}^1, \ldots, s_{wn}^1\} \tag{5}$$

3.2c *Contextual words*: Consider the keyword s_z which occurs d number of times in the keyframe. From this, the more frequently occurring keywords are identified from the particular keyframe. These frequently occurring keywords are considered as contextual words. The contextual words are expressed by the following equation,

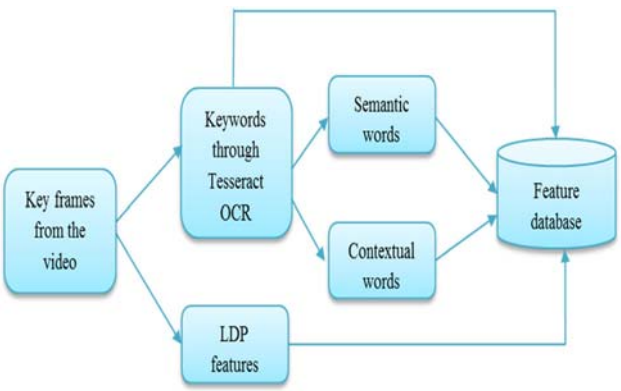

Figure 2. Extraction of features from the keyframes.

$$U = \{u_1, u_2, \ldots, u_q\} \tag{6}$$

where, u_q expresses the contextual words in the keyframes, and U indicates the total contextual words in the keyframe.

3.2d *Local Directional Pattern*: The LDP features signify the direction of the pixels of the individual video frames. The image features are extracted from keyframes by applying 8 different masks. The LDP features [17] are extracted from the keyframe R_i^k . The masks are applied in reference to the center point of the pixel. Consider the keyframe R_i^k has the center pixel as (h_c, k_c) and the LDP features are obtained from the keyframe as

$$LDP(h_c, k_c) = \sum_{f=0}^{7} O(r_f - r_g)2^f \tag{7}$$

Where r_f indicates the f^{th} mask used for obtaining the LDP feature.

3.3 *Clustering the features: FCM algorithm*

The features extracted from the keyframes are stored in the feature database. The next major task in the video retrieval process is the clustering of the features into G number of clusters, which is carried out with the use of the FCM algorithm. The mathematical formulation of the FCM algorithm [27] is as follows,

The initial step in the FCM clustering is the formulation of the fuzzy matrix, by computing the Euclidean distance measure. The fuzzy matrix is expressed as follows,

$$V = \sum_{p=1}^{s} \sum_{e=1}^{l} J_{pe}^r O_{pe}; \ 1 \le r \le \infty \tag{8}$$

where r refers to the fuzziness variable and O_{pe} indicates the Euclidean distance measure and it is measured as,

$$O_{pe} = \|y_p - X_e\| \tag{9}$$

where X_e refers to the cluster center and it is expressed as,

$$X_e = \frac{\sum_{p=1}^{s} J_{pe}^r y_p}{\sum_{p=1}^{s} J_{pe}^r} \tag{10}$$

The cluster center modifies the fuzzy matrix, and it is expressed as,

$$J_{je} = \frac{1}{\sum_{l=1}^{j} \left(\frac{O_{pe}}{O_{le}}\right)^{\frac{2}{r-1}}} \tag{11}$$

FCM executes for the finite interval of time and finds the optimal centroid for the clustering. The centroids calculated through the FCM algorithm are expressed as follows,

$$C = \{C_1, C_2, \ldots, C_j, \ldots, C_G\} \tag{12}$$

where, C_j refers to the j^{th} cluster centroid, and G refers to the total number of clusters.

3.4 *Video retrieval using DBN classifier*

Finally, after clustering the features using the FCM approach, the clustered features are fed to the DBN classifier for video retrieval. DBN [28] gets the clustered features from the FCM and tries to find suitable cluster matching with the query from the user.

The proposed deep learning scheme using DBN performs video retrieval using the training and the testing steps. Figure 3 presents the detailed architecture of video retrieval using the DBN Classifier (A deep learning scheme).

The proposed deep learning system has three layers, namely RBM layer 1, RBM layer 2, and the MLP layer. The centroids of the clusters are fed as training input to the proposed deep learning scheme, the DBN finds the optimal centroid related to the query. The three layers of the DBN are interconnected with each other, with the output of one layer fed to the consecutive layer. The expression for the various layers is briefly described below:

The RBM layer 1 has visible neurons and hidden neurons for the processing. Both the input and the hidden layers of the RBM layer 1 are represented as follows:

$$A^1 = \left\{A_1^1, A_2^1, \ldots, A_j^1, \ldots, A_G^1\right\};\ 1 \le j \le G \tag{13}$$

$$S^1 = \left\{S_1^1, S_2^1, \ldots, S_x^1, \ldots, S_y^1\right\};\ 1 \le x \le y \tag{14}$$

where, A_j^1 and S_x^1 represent the j^{th} visible neuron and x^{th} hidden neuron of the RBM layer 1. The terms G and y refer to the total number of input and hidden neurons in the RBM layer 1. The input layer of the RBM 1 is fed with the features of the centroid of each cluster. The visible and the hidden layer of the RBM contains the biases, which are represented as,

$$a^1 = \left\{a_1^1, a_2^1, \ldots, a_j^1, \ldots, a_G^1\right\} \tag{15}$$

$$b^1 = \left\{b_1^1, b_2^1, \ldots, b_x^1, \ldots, b_y^1\right\} \tag{16}$$

where a_j^1 indicates the bias corresponding to the j^{th} visible layer and b_x^1 corresponds to the bias of the x^{th} hidden layer of RBM 1. The weights between the visible and the hidden neurons are given by the following expression,

$$Z^1 = \left\{Z_{jx}^1\right\};\ 1 \le j \le G;\ 1 \le x \le y \tag{17}$$

where Z_{jx}^1 refers to the weight present amidst the j^{th} visible and x^{th} hidden neuron.

For computing, the output of the RBM layer 1, the bias present in the hidden units, feature inputs, and the weights are used. The following expression indicates the output for the RBM layer 1 obtained through the hidden units,

$$S_x^1 = \sigma\left[b_1^1 + \sum_j A_j^1 Z_{jx}^1\right] \tag{18}$$

where, σ indicates the activation function for computing the output of the RBM 1, which is expressed by the following expression,

$$S^1 = \left\{S_x^1\right\};\ 1 \le x \le y \tag{19}$$

After computing the output at the RBM layer 1, the output is fed as the input to the visible units of the RBM layer 2. Thus, it is necessary to provide a similar number of visible units in RBM 2 as the output layer of RBM 1. The expression for the input of the RBM 2 is represented as follows,

$$A^2 = \{A_1^2, A_2^2, \ldots, A_G^2\} = \left\{S_x^1\right\};\ 1 \le x \le y \tag{20}$$

where S_x^1 indicates the output of the x^{th} layer in RBM 1. The hidden unit present in the RBM 2 is present as follows,

$$S^2 = \left\{S_1^2, S_2^2, \ldots, S_x^2 \ldots, S_y^2\right\};\ 1 \le x \le y \tag{21}$$

Similar to the RBM layer 1, the RBM 2 has biases in both the input and the hidden units. The bias corresponding to the RBM layer 2 is represented as a^2 and b^2 . The weight vector of the RBM layer 2 is given as follows,

$$Z^2 = \left\{Z_{xx}^2\right\};\ 1 \le x \le y \tag{22}$$

where Z_{xx}^2 implies the weight between the x^{th} hidden unit and the x^{th} visible unit of the RBM layer 2. The expression

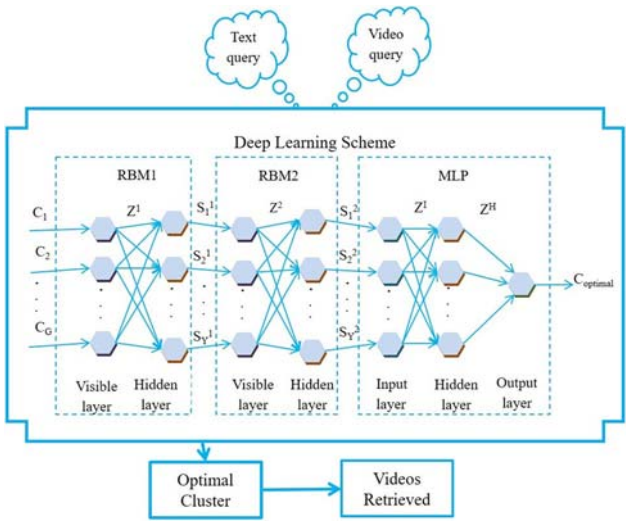

Figure 3. The architecture of video retrieval using DNB classifier.

for the output from the RBM layer 2 is represented as follows,

$$S_x^2 = \sigma\left[b_x^2 + \sum_j A_j^2 W_{xx}^2\right] \forall A_j^2 = S_x^1 \tag{23}$$

where b_x^2 refers to the bias present in the x^{th} hidden unit. The hidden unit of the RBM layer 2 is represented as follows,

$$S^2 = \left\{S_x^2\right\};\ 1 \le x \le y \tag{24}$$

The output of the RBM layer 2 is directly fed to the MLP layer, and thus, the visible units present in the MLP are represented by the following equation,

$$M = \left\{M_1, M_2, \ldots, M_x, \ldots, M_y\right\} = \left\{S_x^2\right\};\ 1 \le x \le y \tag{25}$$

where M_x indicates the x^{th} input unit of the MLP layer. The hidden layer units of the MLP layer are expressed by the following equation,

$$N = \{N_1, N_2, \ldots, N_m, \ldots, N_n\};\ 1 \le m \le n \tag{26}$$

where n indicates the total number of units in the MLP layer. The MLP layer finds the optimal centroid among the input centroids, and thus, the output of the MLP layer comprises only one layer, and it is represented as,

$$C = \left\{C_{optimal}\right\} \tag{27}$$

The output of the MLP layer depends on the weights in the input and the hidden units. The weight vector corresponding to the input unit of the MLP is expressed as,

$$Z^I = \left\{Z_{xm}^I\right\};\ 1 \le x \le y;\ 1 \le m \le n \tag{28}$$

where Z_{xm}^I indicates the weight between the x^{th} input unit and m^{th} hidden unit of the MLP layer. Now, the expression of the hidden unit in the MLP layer is expressed as,

$$N_m = \left[\sum_{x=1}^{y} Z_{xm}^I * M_x\right] Y_m \ \forall b_1 = S_x^2 \tag{29}$$

where Y_m indicates the bias present in the input unit. The expression for the weights present in the hidden unit of the MLP layer is expressed as,

$$Z^H = \left\{Z_m^H\right\};\ 1 \le m \le n \tag{30}$$

The final output of the MLP layer depends on the hidden layer output and the weights present in the hidden unit. The expression for the MLP layer output is expressed as,

$$C_{optimal} = \sum_{m=1}^{n} Z_m^H * Z_m \tag{31}$$

where, Z_m^H is the weight of the hidden unit of the MLP.

3.4a *Training phase*: This section presents the training procedure relating to the DBN. For the training, the features representing the centroids of the clusters are given as the training input to the RBM layer 1. As the proposed deep learning scheme involves the RBM and the MLP layer, each layer is trained using different algorithms. The training procedure identifies suitable weights for the testing process.

I. Training of RBM layers

Initially, the RBM layer is fed with the centroid features for the training. The RBM layer 1 is consecutively connected with the RBM layer 2 and hence, the RBM 1's output serves as the input for the RBM 2. The training procedure aims to identify the optimal weights for both the RBM layers. The training of the RBM layer is bound by the existing backpropagation algorithm.

II. Training of MLP

After training the RBM layers, the input of the MLP layers is taken from the output of RBM layers. In this work, the MLP layer is trained by the existing gradient descent algorithm which provides the expression for the weight update. After training, the optimal weight is identified for the input and the hidden layers for the MLP. The training procedure for the MLP layer is defined below:

i) In the initial step, the weights of the input layer Z^I and the hidden layer Z^H are chosen randomly based on expression (28) and (30).
ii) As the input for the MLP layer depends on the output of the RBM layer 2, the input sample $\left\{S_x^2\right\}$ is fed as the training input.
iii) In the next step, the output of the MLP layer, $C_{optimal}$, is found based on the equation (31).
iv) The training procedure finds the weight based on the minimal error, such that the weight providing minimal output error is considered to be the optimal weight. Hence, the average error is computed based on the following expression,

$$E_{avg} = \frac{1}{V}\sum_{i=1}^{V}\left(C_{optimal}^i - T^i\right)^2 \tag{32}$$

where $C_{optimal}^i$ indicates the output of the MLP and T^i indicates the target response.
v) In the next step, the partial derivative of the weights in the input and hidden units of the MLP is computed by the following expression,

$$\Delta Z_{xm}^I = -\eta \frac{\partial E_{avg}}{\partial Z_{xm}^I} \tag{33}$$

$$\Delta Z_m^H = -\eta \frac{\partial E_{avg}}{\partial Z_m^H} \tag{34}$$

where η refers to the learning rate of the gradient descent algorithm.

vi) The weight update for the weights present in the input and hidden units based on the gradient descent algorithm is expressed below:

$$Z_{xm(G)}^I(t+1) = Z_{xm}^I(t) + \Delta Z_{xm}^I \tag{35}$$

$$Z_{m(G)}^H(t+1) = Z_m^H(t) + \Delta Z_m^H \tag{36}$$

where, $Z_{xm}^I(t)$ and $Z_m^H(t)$ indicate the weights present in the input and the hidden unit of the MLP layer.

vii) Based on the newly computed weight, the MLP layer output and the corresponding average error are computed using the equation (32).

viii) The steps are repeated until the optimal weight is found with a minimum average error.

3.4b *Testing phase*: Consider the user query Q provided to the system, the query from the user can be a video query or a text query. After the query is given to the video retrieval system, the features are pulled out from the query and provided as input to the RBM layer 1. The proposed deep learning based video retrieval system analyzes the feature input of the query and identifies the optimal centroid $C_{optimal}$ suitable for the query. After finding the suitable cluster for the query, all the video contents belonging to the optimal centroid $C_{optimal}$ are retrieved by the proposed deep learning based video retrieval system.

4. Results and discussion

This section briefly explains the simulation results achieved by the proposed lecture video retrieval using deep learning scheme. Simulation is done by choosing different query videos and the results are evaluated based on metrics namely recall, precision, and F-measure.

4.1 *Experimental set-up*

The proposed approach is executed on 2GB RAM, Windows 10 OS with Intel core i-3 processor and is implemented on MATLAB.

4.1a *Database description*: Experimentation of the proposed video retrieval using a deep learning scheme is carried out by considering 100 videos of 6 categories. Each category includes 15 video contents and hence most commonly used for video retrieval. The categories include agriculture, India 2020, pollution, quantum optics, etc.

4.1b *Performance metrics*: Three performance metrics are considered for the evaluation of the deep learning technique for the lecture video retrieval. They are recall, precision, and F-measure.

Recall: It is defined as the fraction of related videos retrieved by the proposed classifier to the total count of related videos available in the database.

Precision: It is defined as the fraction of the total count of related videos retrieved by the proposed classifier to the total count of related and not related videos available in the database.

F-measure: F-measure can be defined as the measure of the harmonic mean of precision and recall values.

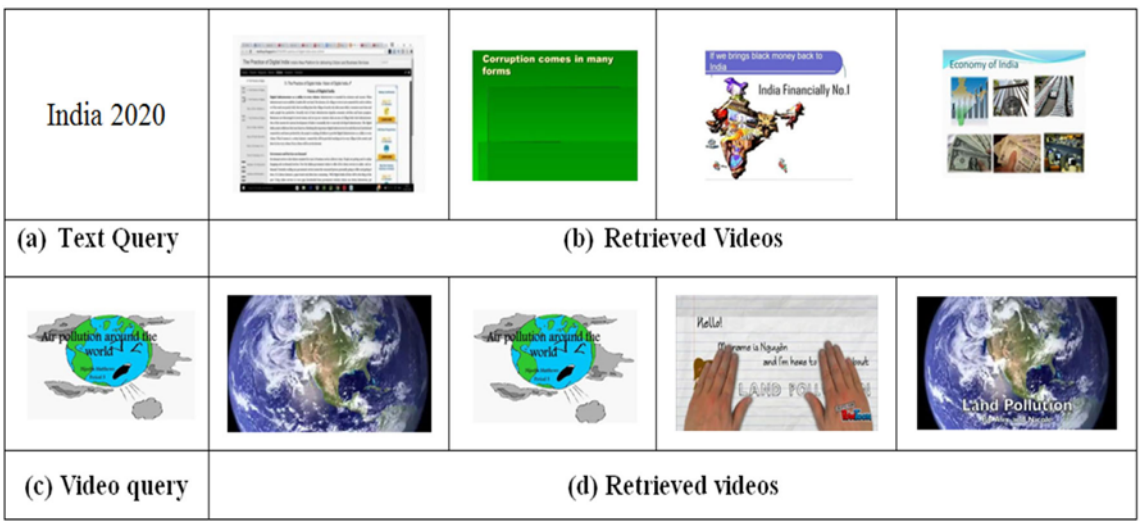

Figure 4. Experimental results of the proposed lecture video retrieval using deep learning scheme for (**a**) text query, (**b**) retrieved videos based on text query, (**c**) video query, and (**d**) retrieved videos based on video query.

4.2 *Experimental results*

The experimental results achieved through the use of the proposed deep learning based video retrieval are presented in Figure 4. Both the video query and the text query have been used for the experimentation.

Figure 4a presents the text query provided by the user, the videos corresponding to the text query retrieved by the proposed scheme are presented in figure 4b. Figure 4c presents the video query provided by the user, and figure 4d presents the retrieved video contents related to video query.

4.3 *Comparative analysis between existing and proposed techniques*

This section performs a comparative analysis between a few existing techniques available for video retrieval and the proposed technique for video retrieval.

Few existing techniques, namely k-NN, and NB have been considered for comparison with the proposed techniques. The existing techniques used for comparison are explained as follows:

- k-Nearest Neighbor (k-NN): The video retrieval is carried out with the help of the k-NN classifier [29].
- Naive Bayes (NB): In this work, the retrieval of the videos has been done through the NB classifier [30].
- Correlated Naive Bayes (CNB): Here, the video retrieval has been done through the CNB classifier [31].
- Semantic Annotation (SA): Video retrieval was done by Annotating videos and incorporating semantic web technologies [32].

A comparative analysis between the proposed method (DBN) and the above discussed existing techniques is done by considering both the user text query and video query. The analysis is done by varying the total number of retrieval and measured based on metrics namely recall, precision, and F-measure.

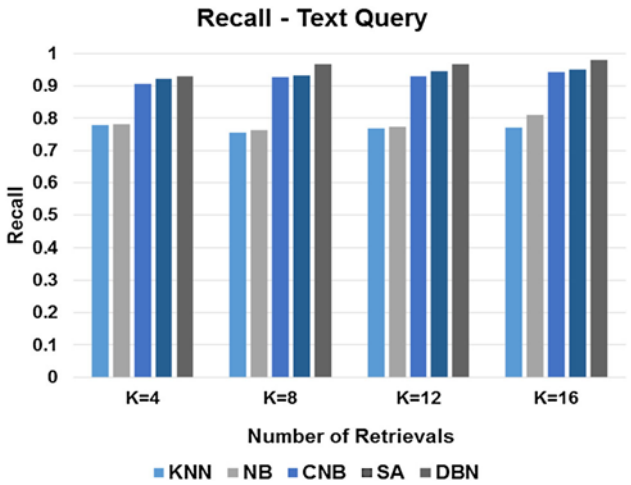

Figure 6. Recall of DBN vs existing techniques.

4.3a *Comparative analysis based on the text query*: A comparative analysis of the proposed deep learning scheme based retrieval scheme with other techniques for various text queries is discussed in this section.

Figure 5 describes a comparative analysis of the proposed deep learning based video retrieval scheme with other techniques based on the precision metric. The existing models, k-NN, NB, CNB, and SA approaches have accomplished precision values of 0.7215, 0.7167, 0.9257, and 0.9322 for the number of retrievals k = 4. Meantime, the DBN method for video retrieval has achieved a high precision with a value of 0.9443.

Figure 6 describes a comparative analysis of the deep learning based video retrieval scheme based on recall metric. The models, namely k-NN, NB, CNB, and SA approaches have achieved the recall values of 0.7864, 0.7826, 0.9282 and 0.9357 for the number of retrievals

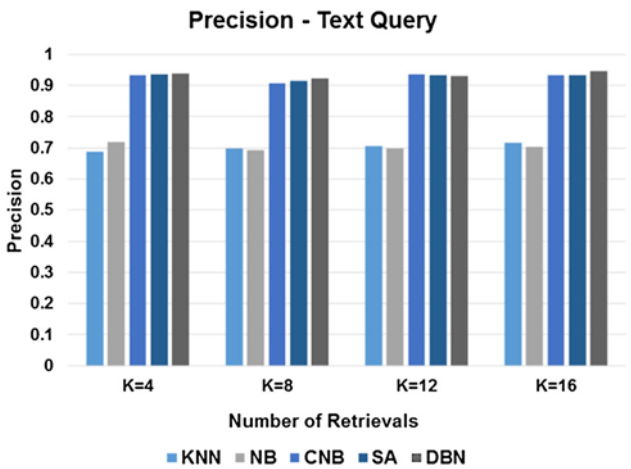

Figure 5. Precision of DBN vs existing techniques.

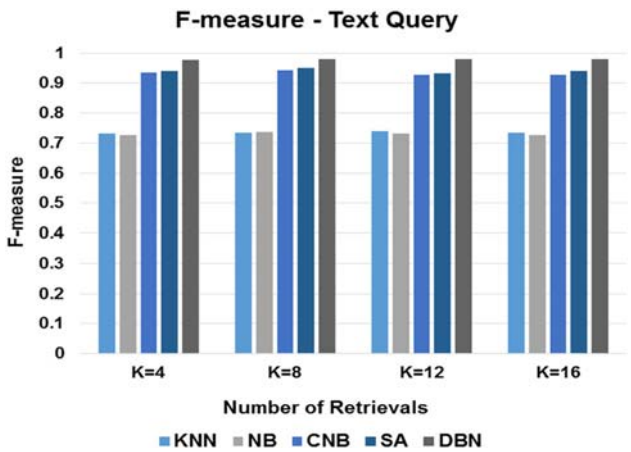

Figure 7. F-measure of DBN vs existing techniques.

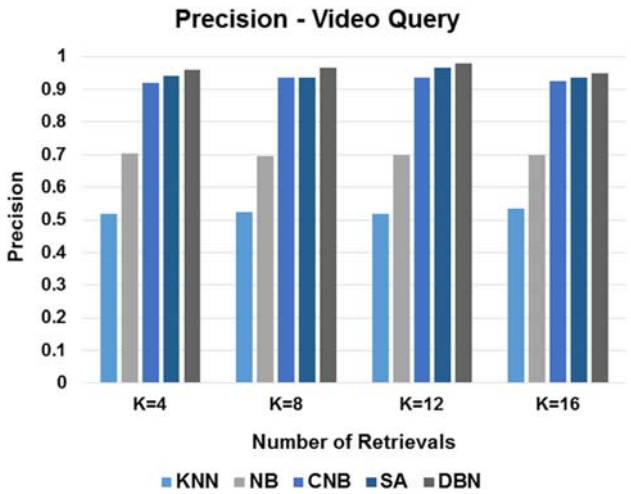

Figure 8. Precision of DBN vs existing techniques.

k = 4, while the proposed DBN method has achieved a high recall with the value of 0.9598.

Figure 7 describes the comparative analysis of the proposed deep learning based video retrieval scheme with other techniques based on the F-measure metric. k-NN, NB, CNB, and SA approaches have accomplished the F-measure value of 0.5899, 0.5985, 0.7444, and 0.7502 for the number of retrievals k = 4. Meanwhile, the proposed DBN scheme has achieved high F-measure with a rate of 0.7933.

4.3b *Comparative analysis based on the video query*: A comparative analysis of the proposed deep learning scheme with various video queries was discussed in this section.

Figure 8 describes the comparative analysis of the proposed deep learning based video retrieval scheme based on a precision metric. The existing models, such as k-NN, NB, CNB, and SA have achieved precision rates of 0.4307, 0.5421, 0.6086, and 0.7213, for K number of retrievals. The proposed DBN scheme has accomplished a high precision with a value of 0.8094.

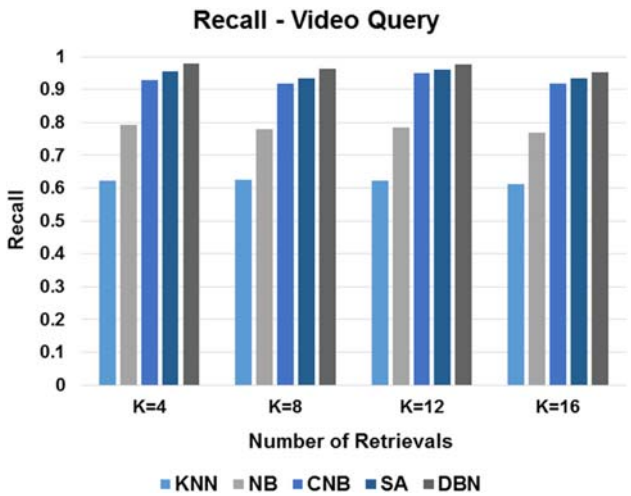

Figure 9. Recall of DBN vs existing techniques.

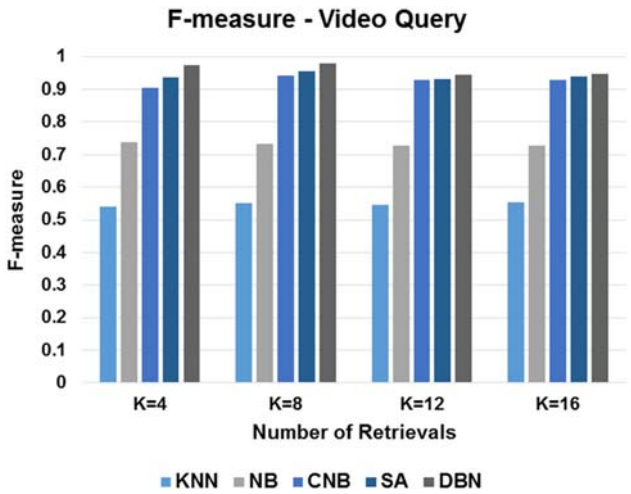

Figure 10. F-measure of DBN vs existing techniques.

Table 1. Comparative analysis of different video retrieval techniques.

Techniques	Precision	Recall	F measure
DBN	0.9620	0.9682	0.9652
SA	0.9370	0.9541	0.9454
CNB	0.9345	0.9420	0.9382
NB	0.7180	0.8096	0.7610
k-NN	0.7156	0.7783	0.7456

Figure 9 illustrates a comparative analysis of the proposed deep learning based video retrieval scheme based on recall metric. The existing models, k-NN, NB, CNB, and SA have achieved the recall values of 0.5302, 0.6505, 0.6578, and 0.7902, for the number of retrieval k. The proposed DBN scheme has achieved high recall with a value of 0.8212.

Figure 10 describes the comparative analysis of the proposed deep learning based video retrieval scheme with other techniques based on the F-measure metric. The results prove that the performance of F-measure is higher for the proposed method when compared to the other metrics.

Finally, the evaluation metrics of the proposed deep learning method (DBN) against other existing techniques for video retrieval is presented in Table 1 for k = 16. The comparative analysis proves that the proposed DBN model outperforms the existing techniques with values of 0.9620, 0.9682, and 0.9652 for recall, precision and F-measure, respectively.

5. Conclusion

A video retrieval strategy using the deep learning scheme was developed. The videos in the video archive have been subjected to the keyframe extraction process, and the necessary keyframes have been extracted. Then, the

features, such as keywords, semantic words, contextual words, and LDP features were extracted from the key-frames and formulated as the database. Then, using the FCM algorithm, the features were clustered into different groups. Then, features representing the cluster centroid were issued to the DBN for training and the DBN found the optimal centroid related to the query. For the experimentation, the research has considered videos from a different category, and both the text query and video query were used for retrieval. The performance of the proposed deep learning scheme for video retrieval was compared with that of the various existing works and evaluated based on metrics, namely, precision, recall, and F-measure. The simulation results prove that the proposed deep learning strategy is more efficient in video retrieval than other retrieval techniques.

Acknowledgements

The authors would like to thank the reviewer(s) and the Associate Editor, Dr. GR Gangadharan, for their constructive comments and suggestions.

References

[1] Yang H, Siebert M, Luhne P, Sack H and Meinel C 2011 Lecture video indexing and analysis using video OCR technology. In: *Seventh International Conference on Signal Image Technology & Internet-Based Systems*. Dijon, 2011, pp. 54–61

[2] Song J, Gao L, Liu L, Zhu X and Sebe N 2018 Quantization-based hashing: a general framework for scalable image and video retrieval. *Pattern Recognit.* 75: 175–187

[3] Hao Y, Mu T, Hong R, Wang M, An N and Goulermas J Y 2017 Stochastic multiview hashing for large-scale near-duplicate video retrieval. *IEEE Trans. Multimed*. 19: 1–14

[4] Fernandez-Beltran R and Pla F 2015 Incremental probabilistic Latent Semantic Analysis for video retrieval. *Image Vision Comput*. 38: 1–12

[5] Fernandez-Beltran R and Pla F 2016 Latent topics-based relevance feedback for video retrieval. *Pattern Recognit.* 51: 72–84

[6] Furini M 2018 On introducing timed tag-clouds in video lectures indexing. *Multimed. Tools Appl*. 77: 967–984

[7] Liang B, Xiao W and Liu X 2012 Design of video retrieval system using MPEG-7 descriptors. *Procedia Eng*. 29: 2578–2582

[8] Kanadje M, Miller Z, Agarwal A, Gaborski R, Zanibbi R and Ludi S 2016 Assisted keyword indexing for lecture videos using unsupervised keyword spotting. *Pattern Recognit. Lett*. 71: 8–15

[9] Roy S, Shivakumara P, Jain N, Khare V, Dutta A, Pal U and Lu T 2018 Rough-fuzzy based scene categorization for text detection and recognition in video. *Pattern Recognit.* 80: 64–82

[10] Memar S, Affendey L S, Mustapha N, Doraisamy S C and Ektefa M 2013 An integrated semantic-based approach in concept based video retrieval. *Multimed. Tools Appl*. 64: 77–95

[11] Furini M and Mirri S 2017 On using on-the-fly students' notes in video lecture indexing. In: *14th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. Las Vegas, NV, pp. 1083–1088

[12] Roy S, Shivakumara P, Jain N, Khare V, Dutta A, Pal U and Lu T 2018 Rough-fuzzy based scene categorization for text detection and recognition in video. *Pattern Recognit.* 80: 64–82

[13] Wu S, Song H, Cheng G and Zhong X 2018 Civil engineering supervision video retrieval method optimization based on spectral clustering and R-tree. *Neural Comput. Appl.* 31: 4513–4525

[14] Chivadshetti P, Sadafale K and Thakare K 2015 Content based video retrieval using integrated feature extraction and personalization of results. In: *International Conference on Information Processing (ICIP)*. Pune, pp. 170–175

[15] Chou C L, Chen H T and Lee S Y 2015 Pattern-based near-duplicate video retrieval and localization on web-scale videos. *IEEE Trans. Multimed*. 17: 382–395

[16] Han X, Singh B, Morariu V I and Davis L S 2017 VRFP: On-the-fly video retrieval using web images and fast fisher vector products. *IEEE Trans. Multimed*. 19: 1583–1595

[17] Chakraborti T, McCane B, Mills S, and Pal U 2018 LOOP Descriptor: local optimal oriented pattern. *IEEE Signal Process. Lett*. 25: 635–639

[18] Yang H, and Meinel C 2014 Content based lecture video retrieval using speech and video text information. *IEEE Trans. Learn. Technol*. 7: 142–154

[19] Li K, Wang J, Wang H and Dai Q 2015 Structuring lecture videos by automatic projection screen localization and analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 37: 1233–1246

[20] Baidya E and Goel S 2014 LectureKhoj: automatic tagging and semantic segmentation of online lecture videos. In: *Proceedings of IEEE International Conference on Contemporary Computing (IC3)*. Noida, pp. 37–43

[21] Nguyen N V, Coustaty M and Ogier J M 2014 Multi-modal and cross-modal for lecture videos retrieval. In: *22nd International Conference on Pattern Recognition*. Stockholm, pp. 2667–2672.

[22] Araujo A and Girod B 2018 Large-scale video retrieval using image queries. *IEEE Trans. Circuits Syst. Video Technol*. 28: 1406–1420

[23] Rahmani F and Zargari F 2018 Temporal feature vector for video analysis and retrieval in high efficiency video coding compressed domain. *Electron. Lett*. 54: 294–295

[24] Lin J, Duan L-Y, Wang S, Bai Y, Lou Y, Chandrasekhar V, Huang T, Kot A and Gao W 2017 HNIP: Compact Deep Invariant Representations for Video Matching, Localization, and Retrieval. *IEEE Trans. Multimed*. 19: 1968–1983

[25] Rouhi A H, and Thom J A 2018 Encoder settings impact on intra-prediction-based descriptors for video retrieval. *J. Vis. Commun. Image R*. 50: 263–269

[26] Patel C, Patel A and Patel D 2012 Optical character recognition by open source OCR tool tesseract: a case study. *Int. J. Comput. Appl.* 55: 975–8887

[27] Bezdek J C, Ehrlich R and Full W 1984 FCM: The fuzzy c-means clustering algorithm. *Comput Geosci*. 10: 191–203

[28] Hinton G E 2009 Deep belief networks. *Scholarpedia*. 4: 5947

[29] Kanungo T, Mount D M, Netanyahu N S, Piatko C D, Silverman R and Wu A Y 2002 An efficient k-means clustering algorithm: analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* 24: 881–892

[30] Balasubramanian V, Doraisamy S G and Kanakarajan N K 2016 A multimodal approach for extracting content descriptive metadata from lecture videos. *J. Intell. Inf. Syst*. 46: 121–145

[31] Poornima N and Saleena B 2018 Multi-modal features and correlation incorporated Naive Bayes classifier for a semantic-enriched lecture video retrieval system. *Imaging Sci. J.* 66: 263–277

[32] Poornima N and Saleena B 2018 Automatic annotation of educational videos for enhancing information retrieval. *Pertanika J. Sci. Technol*. 26: 1571-1590